На правах рукописи

**КАСЬЯНОВ Артем Сергеевич**

# НОВЫЕ МЕТОДЫ ОБРАБОТКИ ДАННЫХ, ПОЛУЧЕННЫХ С ПОМОЩЬЮ СОВРЕМЕННЫХ ТЕХНОЛОГИЙ СЕКВЕНИРОВАНИЯ, ДЛЯ РЕШЕНИЯ ЗАДАЧ АНАЛИЗА ЭКСПРЕССИИ ГЕНОВ

03.01.03 – Молекулярная биология

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата физико-математических наук

Москва 2012

Работа выполнена в Лаборатории биоинформатики и системной биологии Федерального государственного бюджетного учреждения науки Института молекулярной биологии им. В.А. Энгельгардта Российской академии наук.

Научный руководитель: Заведующий Лабораторией биоинформатики и системной биологии Федерального государственного бюджетного учреждения науки Института молекулярной биологии им. В.А. Энгельгардта Российской академии наук, доктор физико-математических наук, профессор Туманян В.Г.

Официальные оппоненты: Профессор факультета биоинженерии и биоинформатики Федерального государственного бюджетного учреждения высшего профессионального образования «Московский государственный университет имени М.В. Ломоносова», кандидат физико-математических наук, доктор биологических наук

Миронов А.А.

Руководитель группы биоинформатики Федерального государственного бюджетного учреждения науки Института общей генетики им. Н.И.Вавилова Российской академии наук, кандидат биологических наук Артамонова И.И.

Ведущая организация: Государственный научный центр Российской Федерации ФГУП Государственный научно-исследовательский институт генетики и селекции промышленных микроорганизмов

Защита состоится «20» декабря 2012 г. в 12[00] часов на заседании Диссертационного Совета Д 002.235.01 при Федеральном государственном бюджетном учреждении науки Институте молекулярной биологии им. В.А. Энгельгардта Российской академии наук по адресу: 119991, г. Москва, ул. Вавилова 34,

С диссертацией можно ознакомиться в библиотеке Федерального государственного бюджетного учреждения науки Института молекулярной биологии им. В.А. Энгельгардта Российской академии наук.

Автореферат разослан «19» ноября 2012 г.

Ученый секретарь диссертационного совета, кандидат химических наук А.М. Крицын

## ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

**Актуальность проблемы.** Последние годы характеризуются бурным развитием технологий высокопроизводительного секвенирования. Одной из главных тенденций является удешевление стоимости секвенирования одного нуклеотида. Увеличение производительности секвенаторов приводит к необходимости разработки более производительного программного обеспечения для обработки данных, полученных с их помощью.

Технологии секвенирования нового поколения наряду со своей основной задачей, т.е. получением последовательностей генома, позволяют решать задачи, связанные с анализом экспрессии генов (RNA-Seq), специфического ДНК-белкового взаимодействия и структуры хроматина (ChIP-Seq), метилирования ДНК (Methyl-Seq). В результате получается, что задачи анализа биологических макромолекул, которые до сих пор решались различными методами (микрочипы, футпринтинг и т.д.) можно решить с помощью технологий секвенирования нового поколения, что является значительным преимуществом, так как оборудование для секвенирования стремительно дешевеет. Специфика применения секвенаторов нового поколения для решения различных биологических задач заключается в методах подготовки образцов и последующей обработке данных с помощью методов биоинформатики. При разработке алгоритмов обработки данных секвенирования нового поколения (next-generation sequencing, NGS) имеют место как чисто алгоритмические сложности, связанные с огромным объемом данных, так и специфические сложности связанные с характером биологической задачи. Хотя методы секвенирования нового поколения возникли совсем недавно, было разработано огромное количество специфичных для них алгоритмов обработки данных; однако вследствие быстро растущих объемов данных и непрерывного развития технологий эффективность существующих алгоритмов недостаточна. Более того, многие существующие алгоритмы были разработаны под решение конкретных задач и неприменимы в других условиях. Таким образом, обработка данных остается лимитирующим фактором, ограничивающим использование технологий секвенирования. В результате разработка новых

алгоритмов остается насущной необходимостью и прогресс в этой области приведет к расширению области применения технологий секвенирования.

В отличии от секвенаторов, построенных на основе классического метода Сенгера, технологии секвенирования нового поколения дают большое количество сравнительно коротких нуклеотидных последовательностей. На данный момент наиболее распространенными технологиями высокопроизводительного секвенирования являются секвенирование путем синтеза с обратимой терминацией (Illumina), пиросеквенирование (Roche), секвенирование путем лигирования (SOLiD), полупроводниковое секвенирование (Ion torrent). Длина чтений (последовательностей, полученных в результате секвенирования), выдаваемых секвенаторами, построенными на основе этих технологий, варьируется от 30 п.н. до 700 п.н., а классический метод Сенгера дает чтения длиной 1000 п.н.. Вследствие этого программное обеспечение, предназначенное для обработки данных, полученных с секвентаров по Сэнгеру, не эффективно при работе с данными нового поколения. Кроме того, вследствие постоянного роста производительности секвенаторов, объем обрабатываемых данных постоянно растет. Данная ситуация осложняется постоянной доработкой существующих технологий, а также появлением абсолютно новых технологических платформ, что, в частности, приводит к изменению шумовых характеристик получаемых данных. Данные трудности постоянно стимулируют создание новых методов обработки NGS данных, потребность в которых на данный момент полностью не удолетворена.

Первичными данными NGS являются относительно короткие фрагменты (чтения, риды) нуклеотидных последовательностей, и основной проблемой при анализе этих данных является их недостаточная длина . Вследствие этого, для обеспечения возможности проведения анализа необходимо иметь достаточно большие покрытия исследуемой биологической последовательности ридами. В результате возникает растет объем данных, получаемых с секвенирующей установки, и возникают вычислительные проблемы работы с такими объемами. Таким образом, алгоритмические подходы, разработанные для традиционных типов секвенирования, не могут применяться для анализа NGS данных независимо от конкретного типа используемого в эксперименте оборудования.

На данный момент разработан ряд подходов для анализа NGS данных, таких как методы, основанные на графах де Брёйна и алгоритмы на основе пробразования Барроуза — Уилера. К сожалению, данные методы основаны на эвристических подходах и являются приближенными, что приводит к получению неоптимальных результатов и потери части информации, содержащейся в исходных данных. Вследствие приведенных выше фактов, задача разработки новых методов и алгоритмов обработки NGS данных остается достаточно актуальной на сегодняшний день.

**Цели и задачи работы.** Целью работы является разработка новых методов и алгоритмов обработки данных, полученных с секвенаторов нового поколения, для решения задачи анализа экспрессии генов. Были поставлены следующие задачи.

1. Разработка методов для анализа дифференциальной экспрессии генов у двух близких, видов на основе данных de novo секвенирования транскриптомов.
2. Аппробация программного обеспечения, разработанного на основе предложенных методов, на базе проекта *de novo* секвенирования транскриптомов двух видов гречихи *F.esculentum* и *F.tataricum*.
3. Разработка методов и алгоритмов для *de novo* сборки транскриптомов полиплоидных организмов.
4. Аппробация программного обеспечения, разработанного на основе предложенных методов, на базе проекта *de novo* секвенирования транскриптома тетраплоида *Capsella bursa-pastoris*.
5. Разработка методов подготовки первичных данных для последующего анализа результатов ChIP-Seq экспериментов для выявления участков ДНК, специфически связывающих белки - регуляторы транскрипции.

**Научная новизна.** Разработаны новые методы для оценки дифференциальной экспрессии генов у двух близких видов на основе данных, полученных с секвенаторов нового поколения. Данные алгоритмы позволили оценить дифференциальную экспрессию транскриптов даже в тех случаях, когда не известен геном изучаемых видов. Публикация предложенного подхода была одной из первых в мире, содержащей способ решения этой задачи.

Предложены новые подходы для первичной обработки ChIP-Seq данных. Предложен новый метод для разделения гаплоидных транскриптов при сборке *de novo* транскриптомов полиплоидов. Разработанные методы позволят повысить эффективность анализа экспрессии генов в RNA-Seq и ДНК-белкового узнавания в ChIP-Seq экспериментах.

**Практическая значимость.** На базе предложенных подходов было разработано программное обеспечение для обработки данных, полученных в результате RNA-Seq и ChIP-Seq экспериментов.

Был разработан вычислительный комплекс для анализа дифференциальной экспрессии генов двух близких видов с использованием данных, полученных с секвенаторов нового поколения. Данное программное обеспечение позволило изучить дифференциальную экспрессию двух видов гречихи *F.esculentum* и *F.tataricum* для которых на данный момент нет опубликованной геномной последовательности.

Был разработан набор программных средств для анализа результатов ChIP-Seq экспериментов. С помощью разработанных инструментов значительно упрощается работа с короткими ридами, полученными с наиболее распространенных на данный момент секвенаторов фирмы Illumina, используемых при анализе структуры сайтов специфического узнавания белков - регуляторов транскрипции.

Был разработан конвейер для сборки *de novo* транскриптома полиплоидного организма с выделением гаплоидных вариатов транскриптов. С использованием данного программного обеспечения была проведена сборка *de novo* и анализ транскриптома тетраплоидного растения *Capsella bursa-pastoris*.

Программы и методы, составляющие материал настоящей диссертации использовались при выполнении работ по ГК от 13 июля 2011 года №07.514.11.4005 «Создание программного комплекса, предназначенного для обработки данных, полученных на секвенирующих установках нового поколения, включая сборку протяженных последовательностей и коррекцию ошибок секвенирования».

**Апробация работы.** Результаты работы были представлены на конференции «Meeting on Advances and Challenges of RNA-Seq Analysis» в городе Халле, Германия в июне 2012 года, на конференции «Bioinformatics of Genome

Regulation and Structure\Systems Biology — BGRS\SB-2012» в г. Новосибирске в июле 2012, на конференции «11th European Conference on Computational Biology» в г. Базель, Швейцария в сентябре 2012,

**Публикация.** Материалы диссертационной работы отражены в 5 публикациях, из них 3 статьи в рецензируемых журналах и 3 публикации в рецензируемых трудах конференций.

**Структура и объем диссертации.** Диссертационная работа состоит из введения, 4 глав, заключения и списка литературы содержащего 104 ссылки. Работа изложена на 104 страницах, содержит 20 рисунков и 17 таблиц.

## СОДЕРЖАНИЕ РАБОТЫ

В **обзоре литературы** рассмотрены основные методы и подходы, используемые для сборки геномных и транскриптомных последовательностей на основе данных, полученных с секвенаторов нового поколения.

Представлен обзор основных технологий секвенирования 2-го и 3-го поколения, таких как технологии Illumina [http://www.illumina.com/pages.ilmn?ID=203], 454/Roche [http://www.454.com/enablingtechnology/the-system.asp], SOLiD [http://marketing.appliedbiosystems.com/images/Product/SolidKnowledge/flash/102207/solid.html], PacBio [www.pacificbiosciences.com], Helicos [www.helicosbio.com], IonTorrent [www.iontorrent.com]. Приведены сравнительные характеристики технологий [Pettersson E. et al., 2009].

Алгоритмы сборки можно разбить на две большие группы: алгоритмы на основе де Брейн графов и алгоритмы на основе «перекрытие-расположение-консенсус» (ПРК, т.н. overlap-layout-consensus, OLC [Myers EW et al., 1995]) подхода. К программам, основанным на алгоритмах на основе де Брейн графов относятся следующие сборщики – EULER [Pevzner P.A. et al., 2001], VELVET [Zerbino D.R. et al., 2008], ABYSS [Simpson J.T. et al., 2009], ALLPATHS [Butler J. et al., 2008], SOAPdenovo [Li R. et al., 2009]. К программам основанным на алгоритмах на основе ПРК подхода относятся следующие сборщики – NEWBLER [Margulies M. et al., 2005], Celera Assembler [Myers EW et al., 2000], ARACHNE [Batzoglou S. et al., 2002], SHORTY [Hossain MS. et al., 2009].

Также в обзоре литературы приведена информация о основных недостатках существующих программ-сборщиков.

## Глава 2. Разработка алгоритмов для анализа дифференциальной экспрессии генов, по данным секвенирования нового поколения транскриптомов близких видов, последовательности полных геномов которых неизвестны.

Оценка дифференциальной экспрессии генов для двух близких видов, последовательности полных геномов которых неизвестны является достаточно важной задачей, возникающей в процессе биологических исследований. Ее решение с использованием технологий секвенирования нового поколения затрудняется тем, что отсутствует возможность сравнения транскриптомов на уровне чтений, вследствие того, что их длина недостаточна для проведения анализа.

Введем понятие «**Сравнение транскриптомов двух видов, последовательности полных геномов которых неизвестны, на основе данных секвенирования**» для этого нам понадобиться ряд дополнительных определений.

$$a = (A|G|T|C)^{*} \quad (2.1)$$

$$A = \{a_1, a_2, a_3, \ldots, a_n\} \quad (2.2)$$

Строки, состоящие из четырех символов, будем обозначать строчными латинскими буквами (2.1), а множество строк прописными латинскими буквами (2.2).

Введем понятие **(m,i,d)-эквивалентности** двух строк. Две строки считаются **(m,i,d)-эквивалентными** в случае если одну строку можно преобразовать в другую с введением не более m замен символов, вставкой не более i символов и удалением не более d символов. Обозначать это отношение будем символом $\overset{m,i,d}{\sim}$.

Введем понятие **(m,i,d)-эквивалентности** двух множеств. Множество A считается **(m,i,d)-эквивалентным** множеству B в случае если для каждой строки из множества A можно найти **(m,i,d)-эквивалентную** в множестве B.

Данное отношение не является коммутативным, в отличие от **(m,i,d)-эквивалентностн** двух строк.

Транскриптом можно считать множеством строк, соответственно процесс секвенирования можно представить в виде операции преобразующей множество строк, соответствующее транскриптому в множество (m,i,d)-эквивалентное множеству подстрок строк принадлежащих множеству, соответствующему транскриптому. Будем обозначать такое отображение символом SEQ.

Пусть $S_A$ - это множество задаваемое следующим образом:

$$S_A = SEQ(A) \tag{2.3}$$

Также введем множества $E_{A,B}$; $U_A$; $U_B$ следующим образом:

$$E_{A,B} = \{a,b \mid a \in A, b \in B, a \overset{m,i,d}{\sim} b\} \tag{2.4}$$

$$U_A = A \setminus E_{A,B} \tag{2.5}$$

$$U_B = B \setminus E_{A,B} \tag{2.6}$$

Введем алгоритмы equal, unique следующим образом Алгоритм equal, принимая на вход множества $S_A$ и $S_B$, конструирует множество $E_{A,B}$ Алгоритм unique, принимая на вход множества $S_A$ и $S_B$, конструирует множество $U_A$.

Задача «*Сравнение транскриптомов двух видов, последовательности полных геномов которых неизвестны, на основе данных секвенирования*», для множеств A и B состоит в нахождении $E_{A,B}$, $U_A$, $U_B$ при наличии только множеств $S_A$ и $S_B$.

Для решения данной задачи требуется построение алгоритмов equal и unique.

Для большей части отображений SEQ (сюда относятся и большинство практически важных случаев) не удастся построить точное решение задачи «*Сравнение транскриптомов двух видов, последовательности полных геномов которых неизвестны, на основе данных секвенирования*», так как получающиеся в результате их применения к исходному множеству множества не содержат необходимой информации. Поэтому имеет большое практическое значение отыскание приближенных решений этой задачи $\tilde{E}_{A,B}$, $\tilde{U}_A$, $\tilde{U}_B$ таких что для них существуют подмножества, являющиеся подмножествами точных решений:

$$\exists C : C \subseteq \tilde{E}_{A,B}, C \overset{m,i,d}{\sim} E_{A,B} \tag{2,9}$$

$$\exists D : D \subseteq \tilde{U}_A, D \overset{m,i,d}{\sim} U_A \quad (2.10)$$

$$\exists E : E \subseteq \tilde{U}_B, E \overset{m,i,d}{\sim} U_B \quad (2.11)$$

В настоящей работе был предложен новый алгоритм для приближенного решения задачи «*Сравнение транскриптомов двух видов, последовательности полных геномов которых неизвестны, на основе данных секвенирования*». Алгоритм состоит из двух этапов – на первом этапе производится сборка чтений в более протяженные последовательности –контиги, на втором этапе полученные два множества контигов выравниваются друг относительно друга с использованием алгоритма парного выравнивания и далее на основе анализа, получившихся выравниваний выделяется набор сходных последовательностей и уникальных для каждого набора данных.

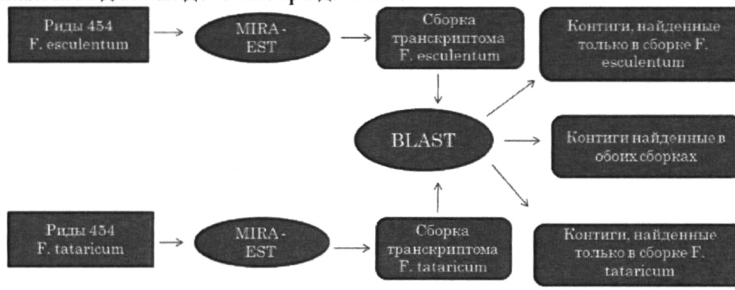


Рисунок 1. Программный комплекс для анализа дифференциальной экспрессии генов двух видов

На основе разработанного алгоритма был разработан программный комплекс (рисунок 1) для анализа дифференциальной экспрессии двух видов, последовательности полных геномов, которых неизвестны, на основе результатов секвенирования их транскриптомов с использованием секвенаторов нового поколения:

1. На первом этапе производится сборка чтений, полученных в результате секвенирования транскриптомов, с использованием программы MIRA [Chevreux B. et al., 1999]. В результате получаем две сборки транскриптомов.
2. Полученные на предыдущем этапе сборки картируются друг на друга с использованием программы BLAST [Altschul S. et al., 1990].
3. Из полученных на предыдущем этапе картирований получаем три набора данных:

- Контиги, найденные только в сборке транскриптома первого организма.
- Контиги, найденные только в сборке транскриптома второго организма.
- Контиги, найденные в обеих сборках.

Описанный пайплайн был использован для сравнительного анализа транскриптомов двух близких видов гречихи *F.esculentum* и *F.tataricum*. Результаты данного исследования были опубликованы в работе [1].

**Секвенирование и сборка транскриптомов двух видов гречихи.** Для секвенирования использовались нормализованные библиотеки кДНК растений *F.esculentum* и *F.tataricum*. Полученные образцы были секвенированы с использованием Roche GS FLX. Было получено 266782 ридов для *F.esculentum* и 229031 ридов для *F.tataricum*. Средняя длина ридов – 349 п.н. для *F.esculentum* и 341 п.н для *F.tataricum*.

Для удаления полиА последовательностей использовался программный пакет SeqClean [http://compbio.dfci.harvard.edu/tgi/software/]. К сожалению данный программный пакет показал полную несостоятельность при применении его для удаления адаптерных последовательностей, поэтому для этих целей автором был разработан новый алгоритм, основанный на получении парных выравниваний последовательности чтения и адаптерной последовательности с пределом по количеству несовпадений (замен, вставок, делеций) в 4 символа и на основе него была разработана специализированная программа для удаления адаптерных последовательностей.. На следующем этапе для анализа предобработанных данных применялся разработанный программный комплекс. Результаты первого этапа его выполнения приведены в таблице 1.

Таблица 1. Характеристика исходных ридов и контигов.

| | Fagopyrum esculentum | Fagopyrum tataricum |
|---|---|---|
| 1 | 2 | 3 |
| Число ридов | 266782 | 229031 |
| Средняя длина(мин-макс) | 349(40-971) | 340(40-976) |
| Число контигов | 25435 | 25401 |

Продолжение таблицы 1.

| 1 | 2 | 3 |
|---|---|---|
| Средняя длина (мин-макс) | 698,4(42-3607) | 703(46-3298) |
| Число ридов на контиг, среднее(мин-макс) | 8,2(2-224) | 7,5(2-295) |

**Аннотация собранных транскрнптомов.** Результаты первого этапа работы программного комплекса были использованы для проведения аннотации. Белковые последовательности, соответствующие контигам были выравнены с неизбыточной белковой базой данных (nr) с порогом на E-value $10^{-6}$. две трети из них имели существенное совпадение. Наиболее значительное совпадение наблюдалось с *Vitis vinifera* [Jaillon O. et al., 2007], *Populus trichocarpa* [Tuskan G.A. et al., 2006], *Ricinus communis* [http://gsc.jcvi.org/projects/msc/ricinus_communis/].

Был проведен автоматический поиск открытых рамок чтения (ORF – open read frame) с помощью The ORF Predictor [Min X.J. et al., 2005]. Предсказано, что в 98% контигов в обоих видах гречихи были найдены открытые рамки считывания длиной более 90 п.н.

Для проведения функциональной аннотации использовалась BLAST2Go. Для обоих видов было аннотировано 60% контигов. Контиги, не имеющие значительного количества BLASTX хитов, но имеющие ORF длиной более 90 п.н., были выравнены с последовательностями из базы данных Pfam. Значительное совпадение было найдено для 1795 последовательностей *F.esculentum* и 1775 последовательностей *F.tataricum*.

**Сравнительный анализ транскриптомов гречихи.** В соответствии с данными сборки и аннотации можно сделать вывод, что присутствует сильная корреляция между количеством ридов, составляющих контиги ортологичных генов *F.esculentum* и *F.tataricum*. Был поставлен вопрос, состоят ли оба транскриптома полностью из ортологичных генов или есть гены, которые экспрессируются у одного вида, а у другого вида они либо отсутствуют, либо не экспрессируются.

В качестве предполагаемых дифференциально экспрессируемых генов (ПДЭГ) были взяты контиги из первого и второго наборов данных, являющихся результатами выполнения рассматриваемого в данном разделе программного комплекса. Всего было найдено 4245 ПДЭГ для *F.esculentum* и 4255 ПДЭГ для *F.tataricum* из них только для 1132 контигов *F.esculentum* и 1588 контигов *F.tataricum* не было найдено совпадений в GenBank[http://www.ncbi.nlm.nih.gov/genbank/] и GeneOntologies[http://www.geneontology.org/] аннотации. Распределение GO категорий в ПДЭГ множествах, соответствующих обоим видам, сильно похоже и только небольшое их количество уникально для двоих видов. Дальнейший анализ показал, что среди уникальных генов присутствуют гены ретротранспозонов и гены, участвующие в биосинтезе сахаров.

Таким образом, с помощью разработанной техники удалось провести крупномасштабный и детальный анализ дифференциальной экспрессии *F.esculentum* и *F.tataricum*, что позволяет говорить о эффективности предложенного подхода.

## Глава 3. Разработка методов подготовки первичных данных для последующего анализа результатов ChIP-Seq экспериментов для выявления участков ДНК, специфически связывающих белки - регуляторы транскрипции.

Большое значение для анализа экспресии генов имеет определение участков ДНК, специфически связывающих белки – регуляторы транскрипции. Знание этих участков позволяет более точно понять каким образом производится регуляция экспрессии генов. Оценка особенностей соответствующих областей ДНК позволяет более точно выявить гены-мишени регуляторных белков на уровне транскрипции.

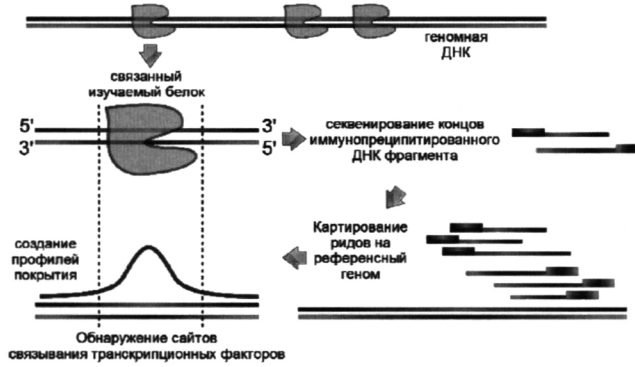


Рисунок 2 Структура ChiP-Seq эксперимента

Процесс обработки данных ChiP-Seq эксперимента состоит из следующих основных этапов:

1) Картирование чтений на референсный геном.
2) Построений профилей покрытия.
3) Нахождение областей связывания транскрипционных факторов.

Второй и третий этап обработки данных достаточно хорошо проработаны и разработано достаточно большое количество программных продуктов решающих эти задачи [ Zambelli, F et al., 2012]. Первый же этап достаточно сильно зависим от типа секвенирующей установки и вследствие постоянных изменений и обновлений вносимых в секвенирующее оборудование на данном этапе возникает потребность в разработке дополнительных алгоритмов предобработки данных и анализа качества картирований.

Был разработан набор алгоритмов и программных средств для предобработки первичных данных ChiP-Seq эксперимента. Разработанное программное обеспечение можно разделить на две категории:

1) Программное обеспечение для подготовки чтений и референсной последовательности для парного выравнивания.
2) Программное обеспечение для анализа качества и проведения классификации результатов выполнения парного выравнивания.

Кроме того, разработан модуль для запуска программ для парного выравнивания ридов BWA [Li H. et al., 2009] и Bowtie [Langmead B. Et al.,

2010], полученных с использованием секвенирующих установок фирмы Illumina.

**Программное обеспечение для подготовки чтений и референсной последовательности для парного выравнивания.** Был разработан алгоритм для оценки статистических характеристик набора чтений:

1) N50(взвешинная медианная длина чтений);
2) Максимальная длина чтения;
3) Средняя и медианная длины чтения;
4) Полное число чтений;
5) Среднее качество набора чтений;

Для ряда задач необходимо исключить часть референсной последовательности перед построением парного выравнивания. К подпоследовательностям, для которых необходимо провести маскирование относятся последовательности гемномных повторов и экзоны. Для маскирования используются база данных RepBase [http://www.girinst.org/repbase/], результаты предсказания тандемных повторов программой TandemSWAN [Valentina Boeva et al., 2006], а также существующая аннотация референсной последовательности. Данный программный модуль использовался при выполнении работ [2,3].

**Программное обеспечение для анализа качества и проведения классификации результатов выполнения парного выравнивания.** Для результатов выравнивания программ BWA и Bowtie было разработано программное обеспечение, оценивающее следующие характеристики набора парных выравниваний:

1) Доля ридов, для которых удалось найти парное выравнивание с референсом.
2) Среднее число замен, делеций и вставок.
3) Максимальное число замен, делеций и вставок в одном выравниваний.
4) Минимальное число замен, делеций и вставок в одном выравнивании.

Был также разработан алгоритм, осуществляющий определение границ покрытия референса ридами, и оценка покрытия данной области ридами. Таким образом возникает возможность классификации наборов ридов и их независимого дальнейшего анализа.

Разработанный набор программных средств повысил эффективность и надежность анализа результатов ChiP-Seq экспериментов для выявления участков ДНК, специфически связывающих белки - регуляторы транскрипции.

## Глава 4. Разработка алгоритма выделения паралогичных генов при сборке диплоидных геномов растений из данных секвенирования транскриптома de novo

Развитие технологий секвенирования привело к возможности получения de novo геномных и транскриптомных последовательностей большого количества различных видов живых организмов. Тем не менее, точное восстановление геномных и транскриптомных последовательностей на основе результатов процесса секвенирования все еще представляет значительную сложность. Геномные и транскриптомные последовательности удается восстановить только в виде набора фрагментов.

Наряду со сложностями, возникающими с геномными и транскриптомными последовательностями, содержащими большое количество повторов, особую трудность составляет процесс восстановления геномных и транскриптомных последовательностей полиплоидных организмов.

Проблема восстановления нуклеотидных последовательностей по данным секвенирования нового поколения для полиплоидных организмов на данный момент полностью не решена. Существующие методики и алгоритмы не дают возможность полностью разделить гаплоидные последовательности и в большинстве случаев требуют наличие полногеномной последовательности для исследуемого вида. Таким образом, в результате работы программ сборщиков мы не можем получить отдельные последовательности гаплотипов, а получаем последовательность состоящую из подпоследовательностей, относящихся к разным гаплотипам. Поэтому значительно затрудняется процесс анализа аллельных вариантов различных генов, а так же не удается выделить

последовательности транскриптов паралогичных генов, что необходимо, например, при изучении эволюционных процессов, приводящих к появлению полиплодии.

Рассмотрим проблему определения и разделения последовательностей транскриптов паралогичных генов для видов, для которых неизвестны полногеномные последовательности, из данных секвенирования более формально.

Введем ряд определений. Обозначим длину строки a, как l(a). Две строки a и b считаются **сильно (m,i,d)-эквивалентиымн**, если $m+i+d << \min(l(a),l(b))$. Паралогичными последовательностями будем считать строки которые сильно (m,i,d)-эквивалентны. Множество кортежей, состоящих из строк из множества A, которые являются попарно паралогичными последовательностями. обозначим как $A_P$.

Таким образом, задачу «*Определения и разделения последовательностей транскриптов паралогичных генов для видов для которых неизвестны полногеномные последовательности*» можно задать следующим образом: необходимо определить множество $A_P$ с использованием только множества $S_A$.

Вследствие того, что в большинстве практических случаев не имеется возможности точного восстановления множества $A_P$, имеет смысл говорить о приближенном решении задачи «*Определения и разделения последовательностей транскриптов паралогичных генов для видов для которых неизвестны полногеномные последовательности*» - $A_p$ .

Для приближенного решения задачи «*Определения и разделения последовательностей транскриптов паралогичных генов для видов для которых неизвестны полногеномные последовательности*» мы предложили алгоритм de novo сборки транскриптомных последовательностей, который позволяет разделить гаплоидные транскрипты после проведения сборки с использованием программ-сборщиков. Данный алгоритм был опробован при сборке и анализе результатов секвенирования транскриптома тетраплоида *Capsella bursa pastoris* [5,6].

Для секвенирования было взяты образцы из третьего поколения потомков одного самоопыляемого растения. Было подготовлено две библиотеки кДНК. Первая библиотека получена из цветков и соцветий и была севенирована с

использованием Roche/454 GS FLX и Illumina Hiseq 2000. Вторая библиотека была получена из растений подвергавшимся различным стрессовым воздействиям (холод, сильное освещение и т.д.). Всего было получено около 50 миллионов чтений, полученных с помощью секвенирующей установки фирмы Illumina, и 1 миллион чтений, полученных с помощью секвенирующей установки фирмы Roche. Была проведена сборка транскриптомных последовательностей на основе этих чтений с использованием трех сборщиков – MIRA, velvet и clc genomics workbench [http://www.clcbio.com/index.php?id=1240]. Было получено 26 сборок с числом контигов от 27463 до 243232. Ни в одной из сборок не удалось найти достаточно удачного разделения паралогичных генов на гаплоиды. При анализе сборок была обнаружена следующая закономерность (рис.5) .


Рис. 5. Участок выравнивания ридов с контигом из сборки.

Можно заметить, что контиги содержат набор позиций, при картировании на которые чтений наблюдается однонуклеотидный полиморфизм (SNP, single nucleotide polymorphism), причем данные позиции можно связать с помощью, соответствующих им ридов. Для этой цели был разработан алгоритм (рис.6) и на основе него была создана программа для выделения паралогичных вариантов. В результате работы разработанного программного обеспечения на выходе получаются три набора последовательностей: последовательности, для которых не найдено паралогичного варианта, последовательности для которых невозможно однозначно определить паралогичные варианты и последовательности для которых удается однозначно определить паралогичные

варианты. На базе этого алгоритма был разработан программный комплекс для выделения паралогичных вариантов генов. Схема разработанного программного комплекса приведена на рисунке 7. Первый этап, не приведенный на схеме 6, представляет из себя получение референсной сборки из тех же ридов, далее проводится выравнивание ридов с полученной «первичной» сборкой. На базе полученных на предыдущем этапе выравниваний проводится поиск SNP позиций. На следующем этапе, производится попытка соединить полученные позиции с помощью помощью парных ридов Illumina и ридов 454. В результате получаем три набора данных. Для набора для которого удается однозначно провести выделение паралогов производится дополнительная досборка.

Первичной сборкой для *Capsella bursa-pastoris* была выбрана сборка, содержащая 27463 контигов. Данная сборка была получена путем установки параметров сборщика таким образом, чтобы паралогичные варианты объединялись (устанавливался максимально возможный размер областей де Брейн графа, которые представляют собой два альтернативных пути, исходящих из одной вершины и входящих в одну вершину, для которых не применяется их разрешение). Результаты обработки данных с использованием, разработанного алгоритма приведены в таблице 3.

На следующем этапе было проведено предсказание ORF с помощью ORF Finder. Только для трех пар паралогичных вариантов не было найдено ORF. Средняя длина предсказанных ORF 580 п.н.. Далее была оценена идентичность паралогичных вариантов внутри каждой пары, как в нуклеотидном пространстве, так и в белковом. Идентичность нуклеотидных последовательностей варьировалась от 79% до 99%. Идентичность белковых последовательностей составила больше 70%.

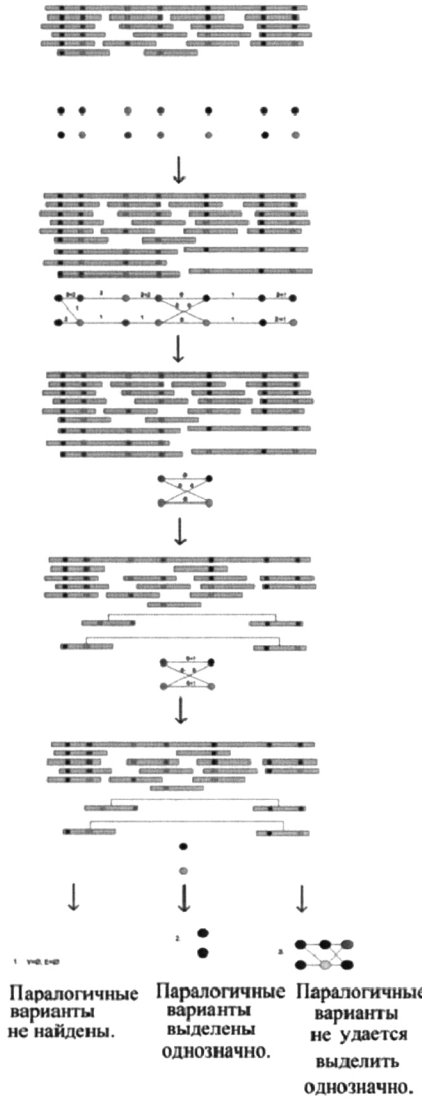


Рис. 6 Алгоритм выделения паралогичных последовательностей

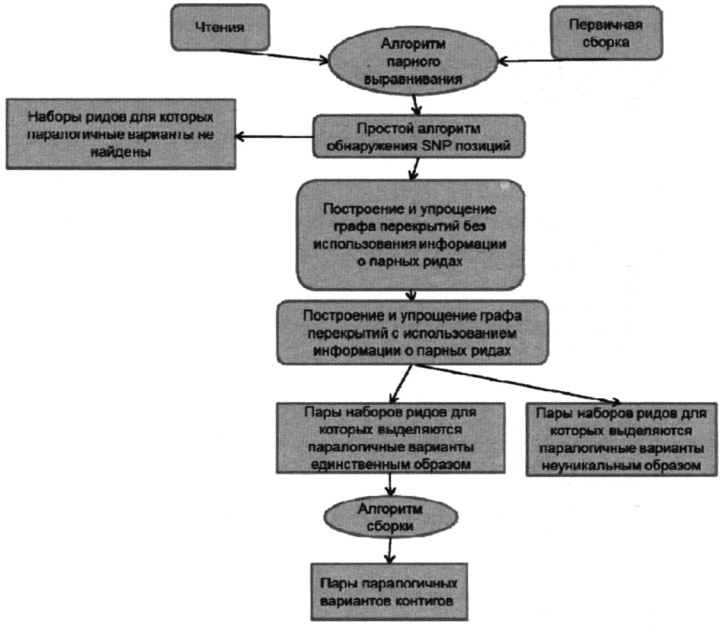


Рис. 7. Программный комплекс для выделения паралогичных вариантов генов при сборке de novo транскриптомных последовательностей.

Таблица 3. Результат обработки данных секвенирования транскриптома *Capsella bursa-pastoris*

| Тип набора | Количество элементов в наборе |
|---|---|
| Последовательности для которых не обнаружены паралогичные варианты | 4428 |
| Последовательности для которых могут быть выделены паралогичные варианты однозначно | 6281 |
| Последовательности для которых не могут однозначно быть выделены паралогичные варианты | 12157 |

## ВЫВОДЫ

1. Предложен метод анализа дифференциальной экспрессии генов, по данным секвенирования транскриптомов близких видов, последовательности полных геномов которых неизвестны. Метод применен для анализа транскриптомов двух видов гречихи *F.esculentum* и *F.tataricum*. Было найдено больше 4200 генов, с потенциально дифференциальной экспрессией, для *F.esculentum* и более 4200 генов, потенциально имеющих дифференциальную экспрессию для *F.tataricum*.

2.Разработан набор средств для предварительной обработки первичных данных при анализе результатов ChIP-Seq экспериментов. Были разработаны средства для подготовки данных, полученных с секвенирующей установки к проведению парного выравнивания с референсной последовательностью, запуска алгоритма парного выравнивания и анализа качества, выполненного выравнивания. Разработано программное обеспечение для маскирования областей геномной последовательности, соответствующей повторам и экзонам. Разработан набор средств для выделения областей покрытых ридами и определения покрытия таких областей.

3. Разработан оригинальный алгоритм для выделения паралогичных вариантов генов при сборке de novo транскриптомов полиплоидных растений. Было проведено севенирование de novo транскриптома тетераплоидного растения *Capsella bursa-pastoris*. В результате однозначно удалось выделить более 6000 пар паралогичных вариантов.

## Список публикаций.

*Статьи в рецензируемых журналах.*

1. Maria D Logacheva, Artem S Kasianov, Dmitriy V Vinogradov, Tagir H Samigullin, Mikhail S Gelfand, Vsevolod J Makeev and Aleksey A Penin De novo sequencing and characterization of floral transcriptome in two species of buckwheat (Fagopyrum)., BMC Genomics 2011, Vol. 12, N30.

2. А.А. Белостоцкий, И.В. Кулаковский, А.С. Касьянов, И.А. Елисеева, В.Ю. Макеев, Характерные предпочтительные расстояния между сайтами

связывания белковых факторов, регулирующих транскрипцию. Биофизика 2011, Т.56, N1, с136-139.

3. I. V. Kulakovskiy, A. A. Belostotsky, A. S. Kasianov, N. G. Esipova, Y. A. Medvedeva, I. A. Eliseeva, and V. J. Makeev A deeper look into transcription regulatory code by preferred pair distance templates for transcription factor binding sites. Bioinformatics 2011, Vol. 27, N19, pp. 2621-2624.

*Тезисы конференций.*

4. Kulakovskiy I, Medvedeva Y, Kasianov A, Vorontsov I, Schaefer U, Bajic V, Makeev V. Comprehensive collection of human transcription factor binding sites models (HOCOMOCO). ECCB'12, 2012.

5. A.S. Kasianov, M.D. Logacheva, N.J. Oparina, and A.A. Penin De novo sequencing, assembly and characterization of transcriptome in the tetraploid plant Capsella bursa-pastoris. Advances and Challenges of RNA-Seq Analysis, Halle/Saale 2012, Germany.

6. Kasianov A.S., Logacheva M.D., Oparina N.J., Penin A.A. De novo sequencing, assembly and characterization of transcriptome in tetraploid plant Capsella bursa-pastoris. BGRS\SB'2012, 2012.

Подписано в печать 18.11.12.

Формат А5

Бумага офсетная. Печать цифровая.

Тираж 100экз. Заказ № 2230

Типография ООО "Ай-клуб" (Печатный салон МДМ)

119146, г. Москва, Комсомольский пр-т, д.28

Тел. 8(495)782-88-39

-10 2